Predviđanje visine godišnjih prihoda na temelju popisnih podataka

Valerija Iva Banić,
Petar Bratulić,
Lea Bundalo

Uvodni opis problema

Tema našeg projekta je predviđanje visine godišnjih prihoda na temelju popisnih podataka. Podaci koje ćemo koristiti dolaze iz tzv. Adult Dataset-a. Adult Dataset ili Income Census Dataset je skup podataka u kojem se uz demografske i ekonomske karakteristike pojedinca nalazi i podatak o prihodima. Ispitanici su podijeljeni s obzirom na visinu godišnjih prihoda u američkim dolarima. Prva skupina je ona koja ima godišnje prihode manje od 50 000 dolara, dok je druga ona čiji su prihodi viši od tog iznosa.

Podaci su prikupljeni 1994. godine iz popisne baze podataka (Census database) i postoji više istraživanja i radova iz strojnog učenja na tu temu. Naš tim pokušat će sa istim tim podacima koristeći razne algoritme i metode strojnog učenja što točnije predvidjeti koje osobe zarađuju više od 50 000 dolara godišnje, a koje manje od tog iznosa.

Skup podataka sastoji se od 48 842 instance koje predstavljaju različite ispitane osobe na popisu stanovništva te 14 demografskih i ekonomskih atributa. Podaci su podijeljeni u dva skupa podataka, train data (u kojem se nalazi 32 561 instanca) te test data (u kojem se nalazi 16 281 insatnca podataka). Više o skupu podataka i njegovoj analizi nalazi se u Jupyter bilježnici eksploratorna_analiza.ipynb.

Cilj i hipoteze istraživanja problema

Cilj našeg istraživanja je odrediti utječu li atributi iz skupa podataka na visinu godišnjih prihoda pojedinca. U eksploratornoj analizi utvrdili smo da neki od atributa uvelike utječu na tu ciljnu vrijednost, nadalje nam je cilj ustanoviti koji atributi imaju više, a koji manje utjecaja te kakav im je utjecaj na prihode (koji ih povećavaju, koji smanjuju i što se događa kombinacijom atributa).

Nakon toga, implementirat ćemo neke od algoritama strojnog učenja kako bi za proizvoljnu osobu za koju znamo vrijednosti atributa mogli procijeniti godišnje prihode. Za kraj, uporedit ćemo različite metode i algoritme koje smo koristili te pronaći onu koja što bolje predviđa visinu godišnjih prihoda pojedinca. Uz to, usporedit ćemo naše rezultate s onima pronađenim u literaturi kako bismo vidjeli koliko su naši algoritmi dobri u odnosu na druge.

Pregled dosadašnjih istraživanja

U radu [2] su na različitim skupovima podataka uspoređeni algoritmi naivni Bayes i C4.5 decision-tree (stablo odluke) algoritam s njihovim hibridnim algortmom NBTree. NBTree algoritam je pogodan kada postoji mnogo atributa koji su bitni za klasifikaciju, ali nisu nužno nezavisni. U navedenom radu, autori su korištenjem hibridnog algoritma NBTree dobili malo poboljšanje preciznosti, no za taj algoritam im je trebalo značajno manje resursa pa je to velika prednost ovog algoritma (c4.5 induced 2213 nodes, NBTree 137).

Rad [3] također se bazira na usporedbi više različitih algoritama. Također se koristi naivni Bayes algoritam, ali se njegovi rezultati uspoređuju s rezultatima algoritama logističke regresije i algoritma slučajnih šuma. Također, u svim algoritmima su korištene četiri metode za rješavanje problema nebalansiranih podataka kako bi se i u odnosu na to mogli usporediti rezultati. Te četiri metode su: bez balansiranja, Random Oversampling (ROS), Random Undersampling (RUS) te kombinacija posljednje dvije metode.

U ostalim radovima koje smo pronašli se uglavnom pojavljuju već nabrojani algoritmi, uz poneku pojavu algoritma k najbližih susjeda.

Materijali, metodologija i plan istraživanja

Kako je vrijednost koju naši algoritmi trebalu predvidjeti binarna varijabla, gdje 0 označava godišnje prihode manje od 50 000 dolara, a 1 označava godišnje prihode veće od 50 000 dolara, problem koji obrađujemo je klasifikacijski.

Zbog toga ćemo problem pokušati riješiti nekim od klasifikacijskih algoritama i metoda strojnog učenja kao što su logistička regresija, neuronske mreže, K najbližih susjeda, naivni Bayes, slučajne šume. Implementirat ćemo više algoritama kako bi mogli međusobno usporediti rezultate, tj. točnost algoritama i pronaći što bolji.

Kako Python ima implementirane mnoge metode i algoritme za regresiju, klasifikaciju i clustering u paketu sklearn (scikit-learn), mi ćemo odabrati one koji će nam trebati te ćemo ih koristiti u našoj implementaciji.

Kako se radi o klasifikacijskom problemu, uspješnost rezultata naših algoritama računat ćemo kao preciznost (accuracy) – zbroj točno klasificiranih u skupini s prihodima manjim od 50 000 i točno klasificiranih u skupini veće od 50 000 dolara podjeljen s ukupnim brojem primjeraka. Skup podataka koji koristimo već je podijeljen na one koji služe za trening i one koji služe za testiranje modela klasifikacije pa ćemo tako preciznost računati na skupu za testiranje.

Očekivani rezultati predloženog projekta

Problem je često rješavan na ovom skupu podataka te ćemo usporediti dobivene rezultate s poznatima. Planiramo dobiti preciznost približno kao i drugi ili bolju.

Literatura

[1] http://archive.ics.uci.edu/ml/datasets/adult (Zadnje pristupljeno 18.4.2019.) Originalni dataset i opis problema

[2] http://robotics.stanford.edu/~ronnyk/nbtree.pdf (Zadnje pristupljeno 18.4.2019.) Rad u kojem je prvi put korišten promatrani skup podataka.

[3] https://storage.googleapis.com/kaggle-forum-message-attachments/160002/5905/Paper%20on%20Machine%20Learning%20for%20Kaggle.pdf (Zadnje pristupljeno 18.4.2019.)